Elements of Structured Data

Asad Raza Virk

2024-09-18

- Data comes from many sources: sensor measurements, events, text, images, and videos.
- The Internet of Things (IoT) is spewing out streams of information. Much of this data is unstructured: images are a collection of pixels, with each pixel containing RGB (red, green, blue) color information.
- Texts are sequences of words and non-word characters, often organized by sections, subsections, and so on.
- Clickstreams are sequences of actions by a user interacting with an app or a web page.
- A major challenge of data science is to harness this torrent of raw data into actionable information.
- To apply the statistical concepts , unstructured raw data must be processed and manipulated into a structured form.

One of the commonest forms of structured data is a table with rows and columns—as data might emerge from a relational database or be collected for a study

Two basic types of Structured Data

1. Numeric
    1. continuous
       such as wind speed or time duration
    2. discrete

    ```
    such as the count of the occurrence of an event
    ```

2. Categorical (takes only fixed set of values)
    1. Binary
       Binary data is an important special case of categorical data that takes on only one of two values, such as 0/1, yes/no, or true/false
    2. ordinal
       ordinal data in which the categories are ordered; an example of this is a numerical rating (1, 2, 3, 4, or 5)

For the purposes of data analysis and predictive modeling, the data type is important to help determine the type of visual display, data analysis, or statistical model.

Data science software, such as R and Python, uses these data types to improve computational performance. More important, the data type for a variable determines how software will handle computations for that variable.

Key Terms for Data Types

1. **Numeric** Data that are expressed on a numeric scale.
    1. Continuous Data that can take on any value in an interval. (Synonyms: Interval, float, numeric)
    2. Discrete Data that can take on only integer values, such as counts. (Synonyms: integer, count)
    3. Categorical Data that can take on only a specific set of values representing a set of possible categories. (Synonyms: enums, enumerated, factors, nominal)
        1. Binary A special case of categorical data with just two categories of values, e.g., 0/1, true/false. (Synonyms: dichotomous, logical, indicator, boolean)
        2. Ordinal Categorical data that has an explicit ordering. (Synonym: ordered factor)

Key Ideas • Data is typically classified in software by type. • Data types include numeric (continuous, discrete) and categorical (binary, ordinal). • Data typing in software acts as a signal to the software on how to process the data.